УДК 81.139

Применение модели случайных блужданий для описания русскоязычных текстов

Крамаренко А. А.[1*], Филимонов В. В.[1], Живодёров А. А.[1,2],Амиева А. М.[1]

[1] Уральский федеральный университет, институт радиоэлектроники и информационных технологий–РтФ, ул. Мира, 32, Екатеринбург, Россия, 620002
[2]Центральная научная библиотека УрО РАН, Екатеринбург, Россия, 620002

Аннотация. В статье рассматривается проблема анализа и атрибуции текста с помощью модели случайных блужданий. В качестве элемента атрибуции предлагается коэффициент пропорциональности в законе Эйнштейна, условно названный коэффициентом диффузии. В ходе работы было сделано предположение, что коэффициент диффузии связан с формой коммуникации между автором и читателем.

Ключевые слова: модель случайных блужданий, закон Эйнштейна, коэффициент диффузии, атрибуция текста.

Application of the random walk model for description Russian-language texts

Kramarenko A.A.[1], FilimonovV.V.[1], Zhivoderov A.A.[1,2], Amieva A.M.[1]

[1] Ural Federal University, Mira, 32, Ekaterinburg, Russia, 620002
[2]Central scientific library URAN, Ekaterinburg, Russia, 620002

Abstract. The paper considers the problem of analysis is and attribution of text using the random walk model. One of elements is coefficient of proportionality in Einstein law, which is named conventionally diffusion coefficient. We suppose that diffusion coefficient is associated with form of communication between author and reader.

Keywords: the ramdom walk model, Einstein law, coefficient of diffusion, attribution of text.

Введение

В настоящее время предпринято много попыток описать явления языка с помощью математического аппарата. Так, в работе [1] рассматриваются кодификаторы дискретных и аналоговых данных в естественном языке, в работе [2] моделируются статистики текста на естественном языке с помощью однопараметрических вероятностных распределений. Также широко используются цепи Маркова для определения авторства текста.

В работах [3; 4] был использован ещё один подход к анализу текстов: проведён статистический анализ распределения в тексте отдельных букв, пар букв (двоек), а также сочетаний трёх букв (троек). При этом учитывались только гласные буквы, так как предполагается, что структуры, возникающие в тексте, связаны с проговариванием слов автором при написании. Выбор последовательности гласных звуков в речевом акте в качестве объекта исследования связан с тем, что гласные звуки играют большую роль при формировании интонационной и ритмической (паралингвистической) структуры речи, чем согласные [3]. На основе специально созданного на кафедре Полиграфии и веб-дизайна УрФУ корпуса текстов, в состав которого входит семь подкорпусов: художественный (проза, поэзия), научный, административный, публицистический и религиозный, социально-политический, было проведено исследование вероятностей появления отдельных гласных букв. В результате выяснилось, что порядок, в котором расположены гласные, ранжированные по убыванию вероятностей, остаётся неизменным от текста к тексту. Этот порядок предложено называть частотной константой языка.

Были исследованы тексты с применением статистики $\chi 2$ как меры отличия текста, представленного случайной последовательностью букв с учётом вероятностей их появления, от реального текста. Оказалось, что значение $\chi 2$ является атрибутом конкретного текста.

Нами была обоснована применимость модели случайных блужданий и разработан алгоритм использования этой модели для описания текстов (материал готовится к печати). Настоящая работа посвящена применению модели случайных блужданий для описания реальных русскоязычных текстов.

1. Модель случайных блужданий

В нашей модели текст рассматривается как цепочка случайных событий –появлений очередной гласной буквы. Основное допущение модели состоит в том, что процесс полагается полностью случайным, то есть появление новой гласной не зависит от предыдущей. Любое случайное блуждание может быть описано законом Эйнштейна, который для двумерного случая выглядит следующим образом:

$$\bar{R}^2 = 4Dt$$

где R — смещение, D — некий коэффициент пропорциональности, аналогичный коэффициенту диффузии для физической системы, t — время. Коэффициент D для удобства будем называть коэффициентом диффузии, а время t соответствует порядковому номеру буквы от начала текста.

Смещение рассчитывается после определенного числа скачков. С точки зрения физического представления нельзя точно определить траекторию случайного блуждания, так как каждый скачок происходит в случайном направлении. Поэтому возникает необходимость в усреднении смещения по множеству случайных процессов. В нашем случае ему соответствует множество фрагментов исследуемого текста.

В рассматриваемой модели движение происходит в некоторой плоскости, каждой букве соответствует свой вектор. Проекции вектора на ось Ox и Oy рассматриваются как смещение в направлении соответствующей оси. Каждое смещение по оси Ox и Oy вычисляется как сумма предыдущего смещения и длина нового скачка. Длина вектора есть величина, обратно пропорциональная частоте появления буквы. Такое значение длины вектора было выбрано для того, чтобы исключить дрейф в сторону букв, встречающихся чаще, чем другие. Углы для каждой буквы выбраны следующим образом: единичная окружность поделена на девять углов по 40º и повёрнута на 5º по часовой стрелке, чтобы направление вектора не совпало с направлением оси, иначе приращение функции по одной из осей было бы равно нулю. Тогда букве А соответствует угол, равный 35º, букве Е — 75º, букве И — 115º и так далее.

Для отработки методики применения модели было взято произведение Ф. М. Достоевского «Подросток», включающее в себя 376744 гласные буквы. С помощью программ Coder и Lines2 отсекались все согласные буквы, цифры, пробелы и знаки препинания, производилось кодирование и подсчёт гласных букв. Далее по формуле:

$$\omega = \frac{n}{N}$$

рассчитывалась частота появления букв в тексте, здесь n — количество появлений отдельной гласной буквы, N — общее количество букв в тексте. Надо отметить, что в расчёт берутся только девять гласных, поскольку буква ё во многих текстах заменена буквой е. Таблица с данными содержит 376744 записей. Фрагмент таблицы в формате MS Excel представлен на рисунке 1.1.

[2]Программы Coder и Lines были специально написаны для исследования сотрудником ЦНБ УрО РАН Л. Г. Горбичем.

назв	Угол	t	n1	ω	r	x	y	Δx	Δy	(Δr)^2
е	75	1	79665	0,211457	4,729103	1,223982	4,567963	1,223982	4,567963	22,364416
о	155	2	101241	0,268726	3,721259	-3,37261	1,572672	-2,14862	6,140635	42,323983
и	115	3	55222	0,146577	6,822353	-2,88325	6,183152	-5,03188	12,32379	177,19549
а	35	4	68938	0,182984	5,464969	4,47664	3,134577	-0,55524	15,45836	239,26931
о	155	5	101241	0,268726	3,721259	-3,37261	1,572672	-3,92784	17,03104	305,48413
и	115	6	55222	0,146577	6,822353	-2,88325	6,183152	-6,81109	23,21419	585,28952
о	155	7	101241	0,268726	3,721259	-3,37261	1,572672	-10,1837	24,78686	718,09616
о	155	8	101241	0,268726	3,721259	-3,37261	1,572672	-13,5563	26,35953	878,59835
е	75	9	79665	0,211457	4,729103	1,223982	4,567963	-12,3323	30,9275	1108,5961
и	115	10	55222	0,146577	6,822353	-2,88325	6,183152	-15,2156	37,11065	1608,7138
о	155	11	101241	0,268726	3,721259	-3,37261	1,572672	-18,5882	38,68332	1841,9196
о	155	12	101241	0,268726	3,721259	-3,37261	1,572672	-21,9608	40,25599	2102,821
о	155	13	101241	0,268726	3,721259	-3,37261	1,572672	-25,3334	41,82866	2391,4179
а	35	14	68938	0,182984	5,464969	4,47664	3,134577	-20,8568	44,96324	2456,6971
а	35	15	68938	0,182984	5,464969	4,47664	3,134577	-16,3801	48,09782	2581,7082
е	75	16	79665	0,211457	4,729103	1,223982	4,567963	-15,1561	52,66578	3003,3927
а	35	17	68938	0,182984	5,464969	4,47664	3,134577	-10,6795	55,80036	3227,7314
я	355	18	22552	0,05986	16,70557	16,642	-1,45599	5,962509	54,34437	2988,8622
е	75	19	79665	0,211457	4,729103	1,223982	4,567963	7,186491	58,91233	3522,3088
у	195	20	24548	0,065158	15,34724	-14,8243	-3,97216	-7,6378	54,94018	3076,7591
е	75	21	79665	0,211457	4,729103	1,223982	4,567963	-6,41382	59,50814	3582,3558
е	75	22	79665	0,211457	4,729103	1,223982	4,567963	-5,18984	64,0761	4132,6813
я	355	23	22552	0,05986	16,70557	16,642	-1,45599	11,45216	62,62012	4052,431
е	75	24	79665	0,211457	4,729103	1,223982	4,567963	12,67614	67,18808	4674,9226

Рис. 1. Фрагмент таблицы значений

Текст был разделён на 3767 траекторий по 100 значений, проведено усреднение траекторий по каждому значению и построена функция $\Delta\overline{r_1^2}(t)$.

Методом аппроксимации с коэффициентом аппроксимации R2 = 0,999 было получено уравнение: $\Delta r12t$ = 324,79t−786,94.

Коэффициент диффузии был рассчитан как тангенс угла наклона прямой. Для текста Ф. М. Достоевского «Подросток» D = 81,20. Также была построена траектория, соответствующая первой тысяче гласных букв текста Ф. М. Достоевского «Подросток» (рис. 2).

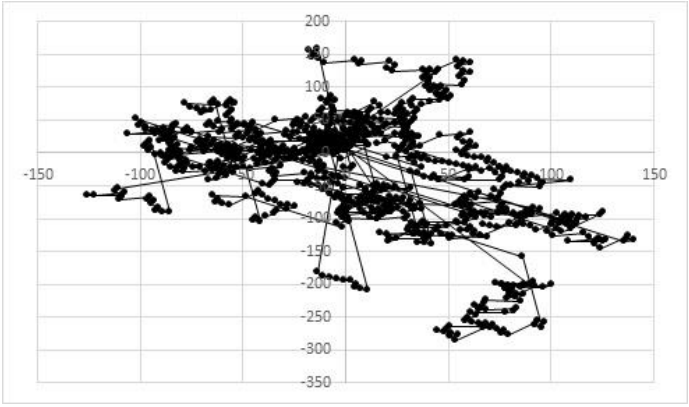

Рис. 2. Траектория, соответствующая первой тысяче гласных букв

2. Применение модели случайных блужданий

На следующем этапе нами были рассчитаны коэффициенты диффузии для двадцати различных текстов. При сравнении результатов была выдвинута гипотеза о зависимости коэффициента диффузии от количества символов в тексте. Чтобы исключить влияние длины текста, необходимо учесть поправку на среднюю флуктуацию, величина которой уменьшается с увеличением выборки, согласно закону больших чисел, пропорционально 1/N0,5. В нашем случае объём выборки соответствует длине текста, а средняя флуктуация определяется среднеквадратичным отклонением (SSD).

$$SSD = B/N^{0,5},$$

где N — длина текста, B — коэффициент пропорциональности.

Для расчёта величины B вне зависимости от авторства и жанровой принадлежности текста с помощью специальной программы Rondo3 были случайным образом сгенерированы две группы текстов, состоящие только из гласных букв: 1) для случая равновероятного появления всех гласных букв; 2) с учётом средних по Корпусу текстов вероятностей появления гласных. Для каждой группы было сгенерировано по десять текстов длиной 5, 10, 20, 50, 100 и 500 тысяч символов (всего 120 текстов).

Для каждого сгенерированного текста был рассчитан коэффициент диффузии и его среднеквадратичное отклонение (SSD). Тексты в рамках групп были объединены в подгруппы по количеству символов, и для каждой из них было вычислено среднее SSD.

Далее методом наименьших квадратов аппроксимирована зависимость среднеквадратичного отклонения от длины текста и найден коэффициент B (рис. 3, 4), с помощью которого можно вычислить SSD для любой длины текста.

[3] Программа Rondo была специально написана для исследования сотрудником ЦНБ УрО РАН Л. Г. Горбичем

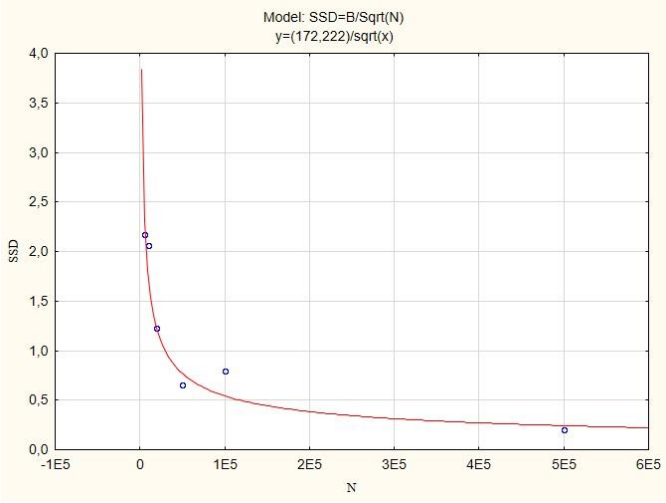

Рис. 3. Зависимость SSD (N) для первой группы текстов

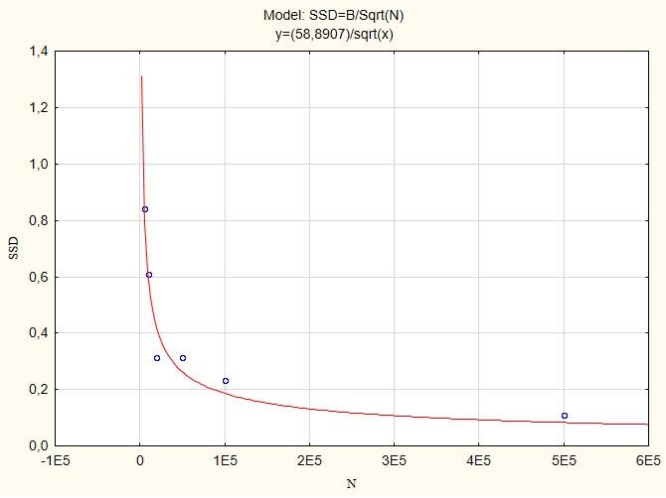

Рис. 4. Зависимость SSD (N) для второй группы текст

Для первой группы текстов B = 58,89 ($p < 0,05$), для второй B = 172,22 ($p < 0,05$). Для расчётов был выбран второй коэффициент, так как, во-первых, тексты с учётом вероятностей появления гласных ближе к реальным текстам, а во-вторых, мы руководствовались тем, что статистически правильнее брать больший доверительный интервал.

Таким образом, была определена величина случайного разброса коэффициента диффузии в зависимости от длины текста. Если коэффициенты диффузии

для разных текстов одной длины отличаются больше, чем определённое нами среднеквадратичное отклонение, то это отличие не случайно и не связано с длиной текста.

После этого было взято по десять текстов из пяти подкорпусов (художественные, научные, административные, публицистические и религиозные тексты), рассчитаны коэффициенты диффузии и SSD. На рис. 5 показано соответствие коэффициента диффузии и номера текста.

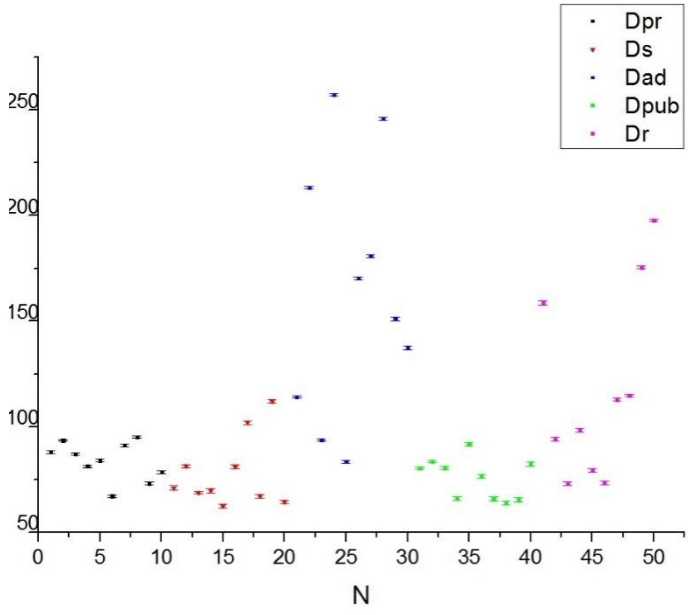

Рис. 5. Коэффициент диффузии в соответствии с номером текста. D_{pr} — художественные тексты (проза), D_s — научные тексты, D_{ad} — административные, D_{pub} — публицистические, D_r — религиозные

Из рисунка 5 видно:

1. Коэффициенты диффузии художественных, научных и публицистических текстов расположены компактно в интервале от 60 до 100.
2. Значения коэффициентов диффузии административных и религиозных текстов имеют сильный разброс в интервалах от 80 до 260 и от 70 до 200 соответственно.
3. Среднее значение диффузии по жанру (выборка небольшая):

Художественные тексты (проза): D = 84,10;
Научные тексты: D = 77,87;
Административные тексты: D = 164,70;
Публицистические тексты: D = 75,52;
Религиозные тексты: D = 114,10.

4. Доверительные интервалы малы, но перекрываются в интервале от 60 до 100, где находятся художественные, научные и публицистические тексты. То есть можно сказать, что эти тексты неотличимы друг от друга по коэффициенту диффузии, и его величина отражает некие общие свойства этих текст

Таким общим свойством, на наш взгляд, может являться «направленность» текстов. Под направленностью мы понимаем реализованную в тексте форму коммуникации между автором и читателем. Таких форм может быть две: субъект-субъектная и субъект-объектная. Первая предполагает партнёрские отношения, в некотором смысле соавторство читателя, совместный поиск смыслов. Вторая — регулирующее воздействие субъекта на объект.

В художественных, научных и публицистических текстах реализуется субъект-субъектная форма коммуникации. В административных и религиозных — субъект-объектная.

3. Результаты и выводы

Значения коэффициента диффузии, соответствующие «субъект-субъектным» текстам, заключены в интервале от 60 до 100. Значения выше 100 характерны для «субъект-объектных» текстов. Среди исследованных текстов есть три, значения коэффициентов диффузии для которых выходят за пределы обозначенных интервалов. Это справочник по русскому языку (D = 111,96), первоначально отнесённый к научному подкорпусу, «Земля и право» (D = 93,66) и «Спутник потребителя» (D = 83,35), первоначально отнесённые к административному подкорпусу.

О справочнике можно сказать, что в нём приведены правила правописания русского языка, а также примеры из различных текстов, составляющие большую его часть. Мы считаем, что причиной выхода текста за указанные пределы послужило наличие фрагментов, принадлежащих большому количеству разных авторов и, как следствие, отсутствие единства стиля.

После более тщательного изучения текстов «Земля и право» и «Спутник потребителя» мы пришли к выводу, что первый следует отнести скорее к научно-популярному подкорпусу, а второй — к публицистическому. То есть оба текста являются «субъект-субъектными».

Таким образом, мы можем сделать предварительный вывод, что с помощью модели случайных блужданий можно разделить тексты субъект-субъектной и субъект-объектной направленности без предварительной экспертной оценки, то есть атрибутировать текст без учёта его смысла.

Список литературы

1. Головко Н. В. Логико-квантитативный аспект теории фиксирования типов языковой информации // Наука, инновации, технологии. 3 (2008). С. 72–79.

2. Закревская Н. С., Ковалевский А. П. Однопараметрические вероятностные модели статистик текста // Сибирский журнал индустриальной математики. 2001. Т. 4. № 2. С. 142–153.
3. Филимонов В. В., Живодеров А. А., Горбич Л. Г., Экспрессия и упорядоченность в письменной речи // В лаборатории учёного. 2012. С. 313–319.
4. Горбич Л. Г., Филимонов В. В., Живодёров А. А. Опыт различения поэтических и прозаических текстов на основе сравнения распределений биграмм гласных букв // Количественные методы в искусствознании: сборник материалов конференции. 2013. С. 163–166.
5. Филимонов В. В., Амиева А. М., Сергеев А. П. Кластеризация русскоязычных текстов с применением статистики χ^2 // Медиатехнологии. 2016. С. 164–174.